# Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 13

Pierre-Luc Germain

ETH Zürich

# Plan for today

- Debriefing on week11 & week12 assignments
- Single-cell epigenomic data (scATAC-seq)
- Chromatin and disease
- Our lines of research
- Evaluation scheme for the course project

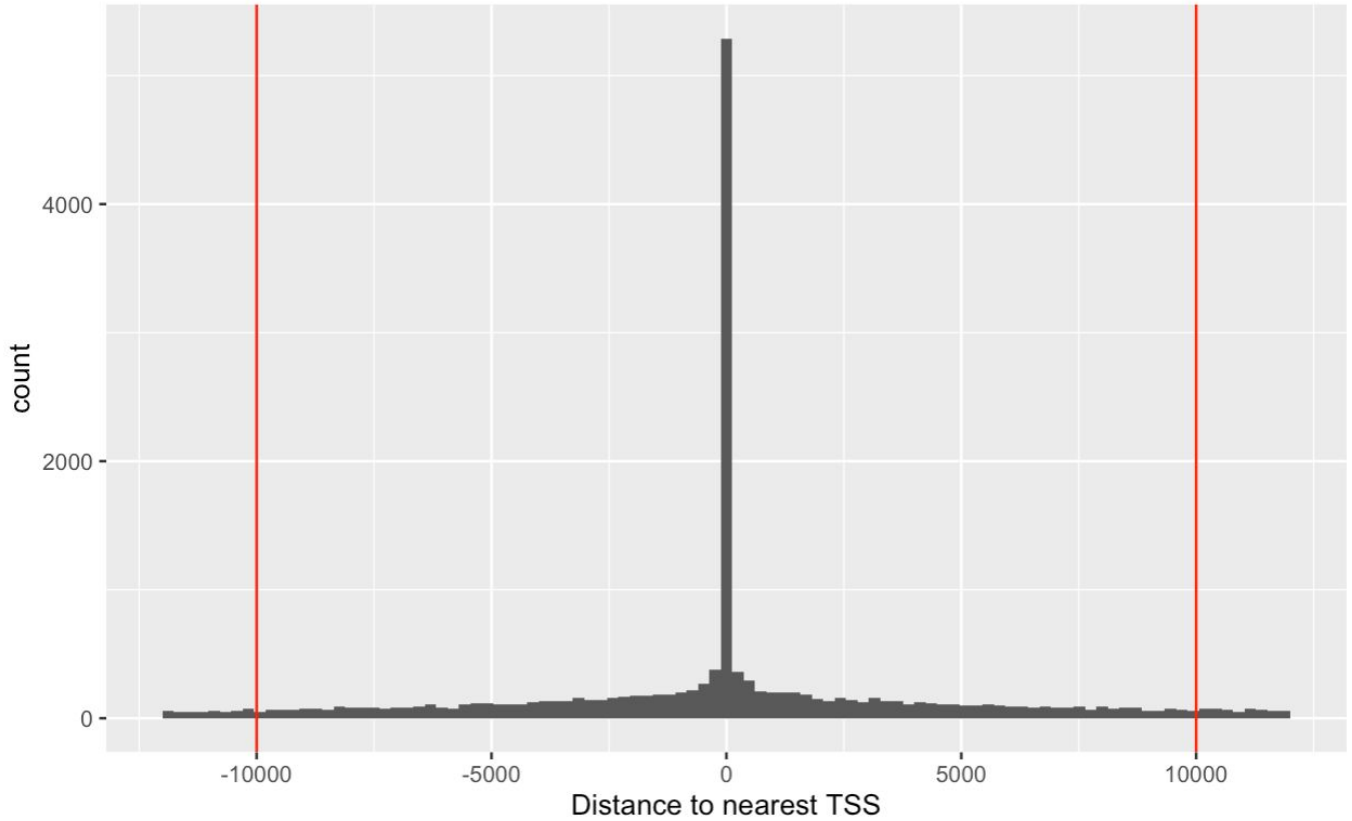
# W11: Debriefing on the assignments

Describe what is tested in the enrichment analysis:

*The enrichment analysis is testing whether the **genes <u>within</u>** the differentially methylated regions (DMRs) are enriched for any Gene Ontology (GO) Biological Process terms compared to **all the genes** in chromosome 22* (Student answer)

1. What we test for enrichment: Genes within DMRs **and not** DMRs
2. What we use as the background: all genes on chromosome 22 **and not** the full genome or the full chromosome 22.

# W12: Debriefing on the assignments



# W12: Debriefing on the assignments

to consider distances up **and** downstream:

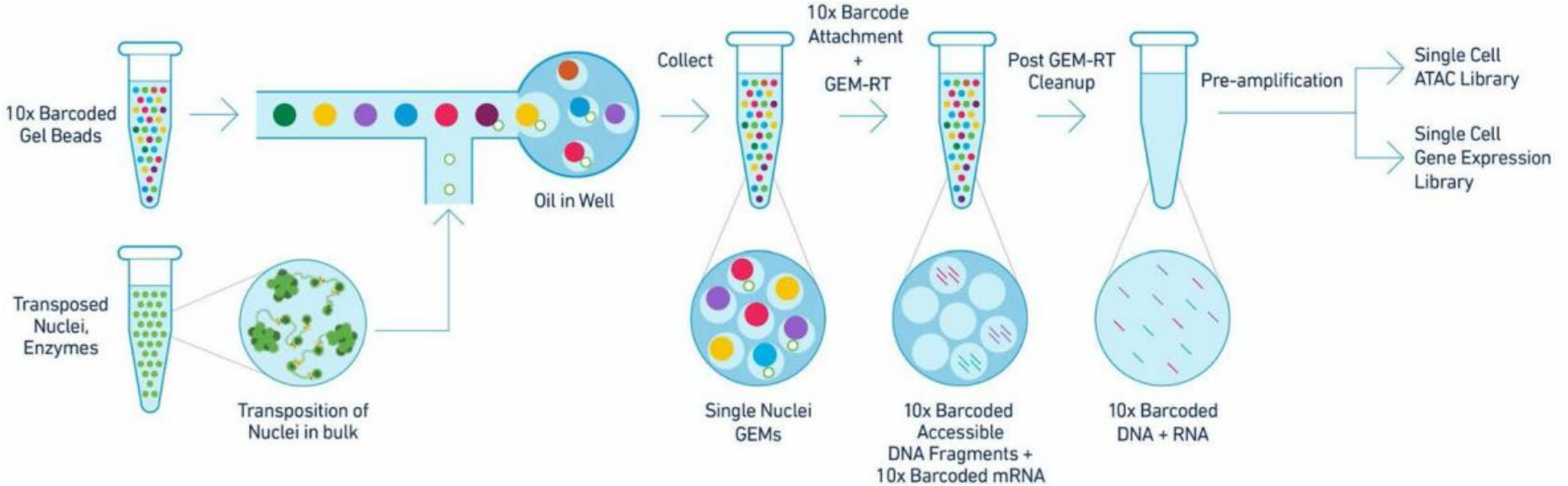
```
peaks2 <- peaks[abs(peaks$distance2nearestTSS) > 10000,]
```

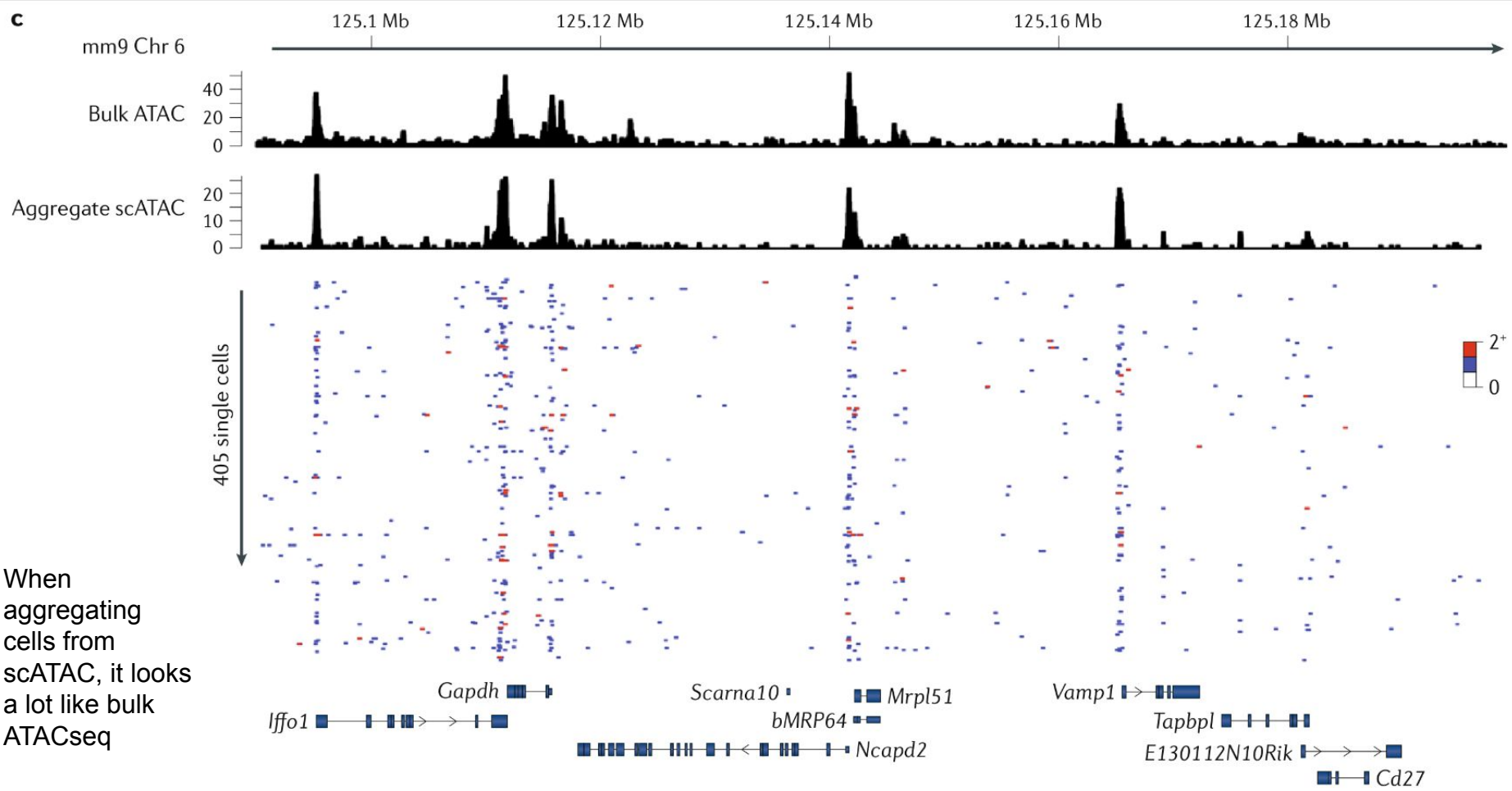
or

```
annotatedPeaks <- epiwraps::annotateRegions(peaks, ensdb,
                                            proximal = c( 2500, 10000, Inf))
```

# Single-cell *-omics

# Single-cell ATAC-seq (and multi-omics)






When aggregating cells from scATAC, it looks a lot like bulk ATACseq

(Mezger et al, Nat Comm 2018)

# Single-cell ATACseq analysis in a nutshell

1. The output of the genome alignment of the data is a “fragment file”, a bed-like file containing the coordinates of each fragment and the associated cell (barcode)

```
#chr    start   end     cell_barcode
chr1    10066   10536   TCAAGCAGTGCGCATC-1
chr1    10073   10278   TCAAGACGTCTGATTG-1
chr1    10073   10305   CGTTCCACAGCGTAGA-1
chr1    10079   10315   TTCAACTTCCGAGAGA-1
chr1    10085   10278   TCGTTCGCATAGGCGA-1
chr1    10091   10303   AGCGTGCTCCCATAGA-1
```

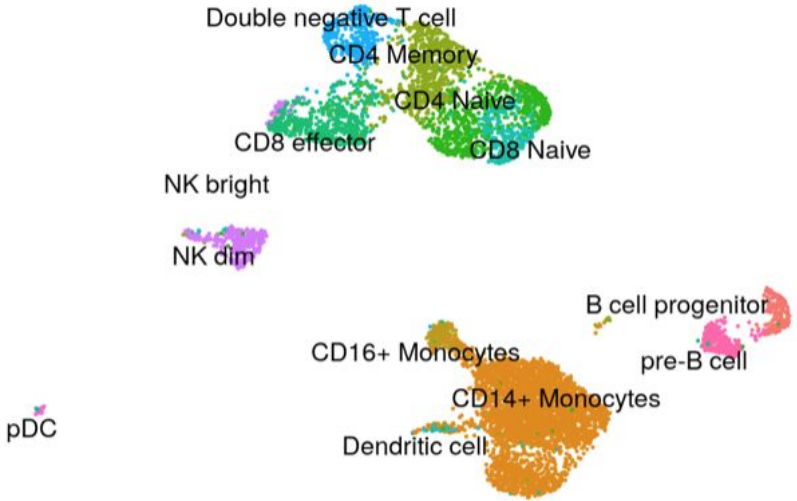
2. From this, we count the number of fragments from each cell overlapping genomic windows (either whole-genome tiles or feature-based)

```
        cell1 cell2 cell3 cell4 cell5 ...
window1     1     0     0     0     0
window2     0     0     0     0     1
window3     0     0     1     0     0
window4     0     0     0     0     0
window5     0     1     0     0     0
...
```

Can be a matrix with hundreds of thousands windows…

(filtering…)

3. Normalization, dimensional reduction (e.g. TF-IDF + LSI), and clustering of the cells:

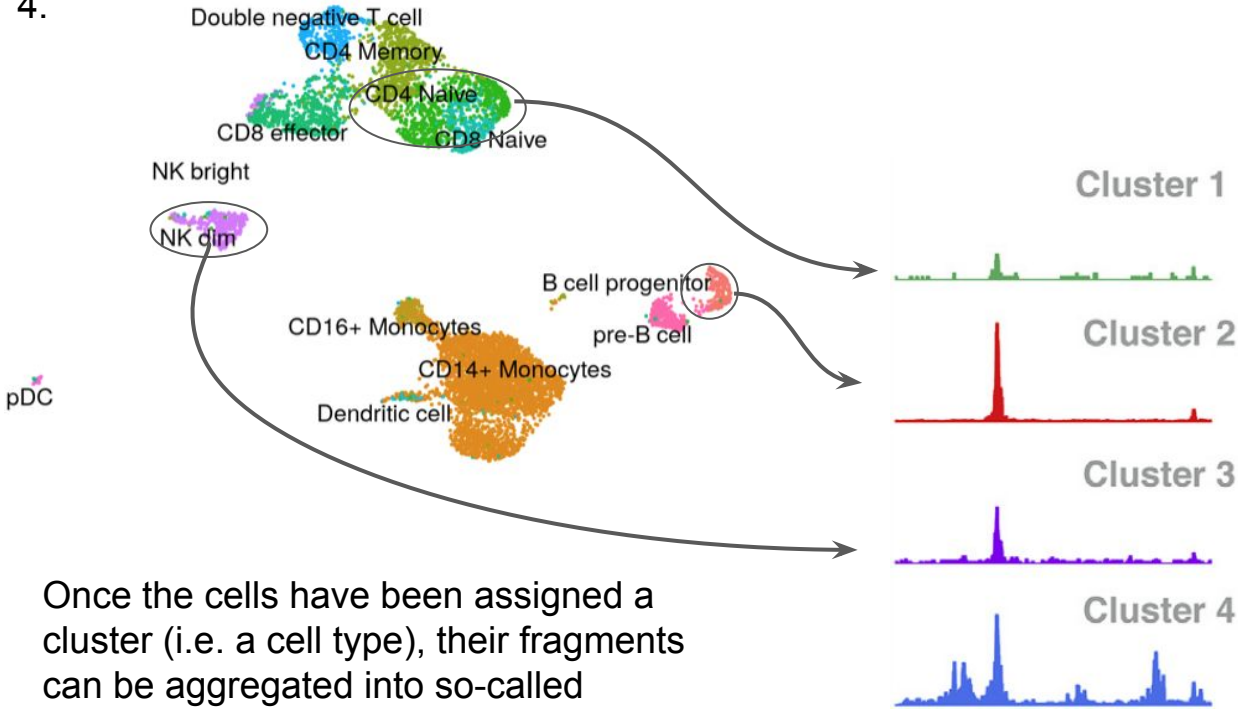



Two main pipelines with excellent documentation:

- ArchR
- Signac

# Single-cell ATACseq analysis in a nutshell

4.



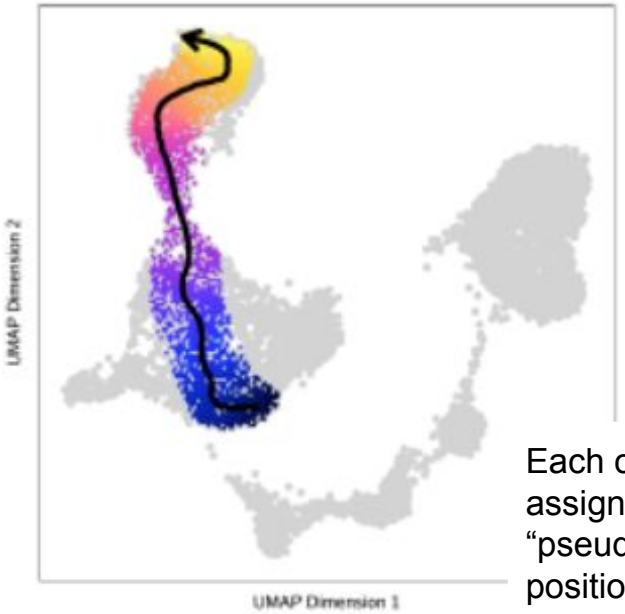

Once the cells have been assigned a cluster (i.e. a cell type), their fragments can be aggregated into so-called "**pseudo-bulk**" profiles

From this point on, the data is pretty much like traditional (bulk) ATACseq data, meaning that you can apply all the tools you're familiar with, but it's cell-type-specific!

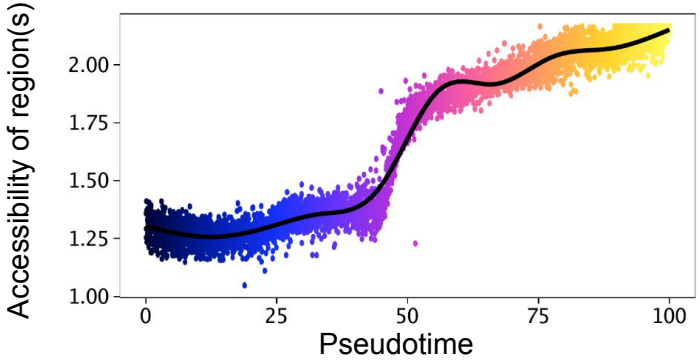
One also often do some work at the pseudo-bulk level (e.g. calling peaks) before going back to the cell-level

# Single-cell ATACseq analysis – doing more at the cell-level

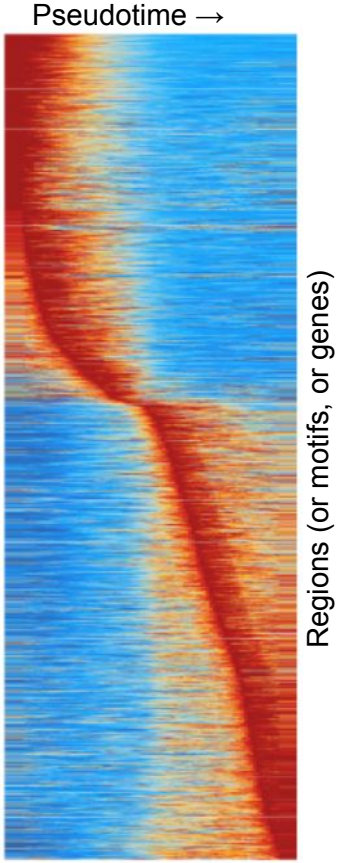
We identify “trajectories” across cells




Each cell can be assigned a “pseudotime”, i.e. it’s position along the trajectory




We can then track the accessibility of regions of interest across this “pseudotime”

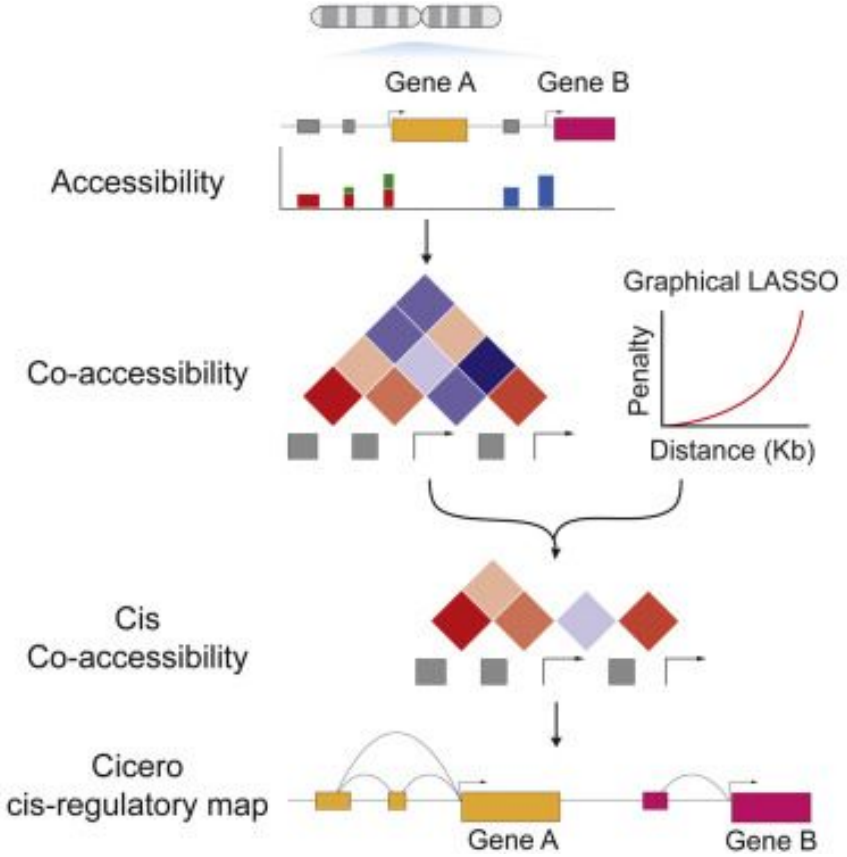



(Adapted from ArchR documentation)

# Single-cell ATACseq analysis – doing more at the cell-level

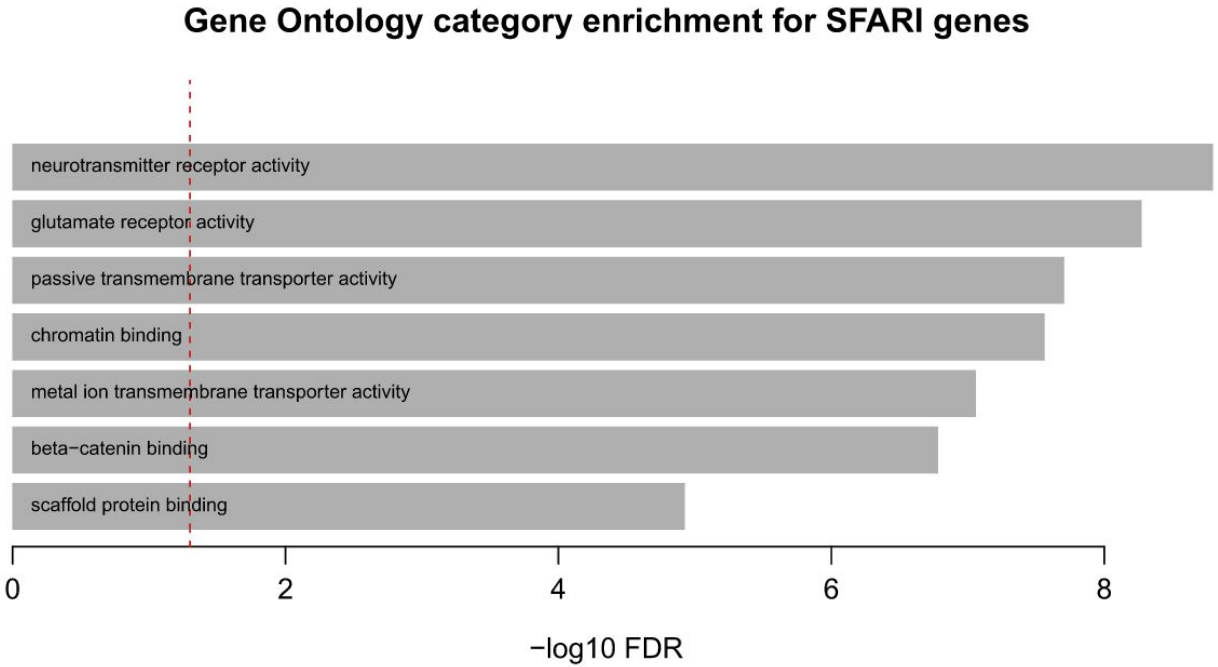
Because we have so many cells, we can use the correlation between the accessibility at distal regulatory elements (i.e. enhancers) and putative TSS to know what genes these regulate

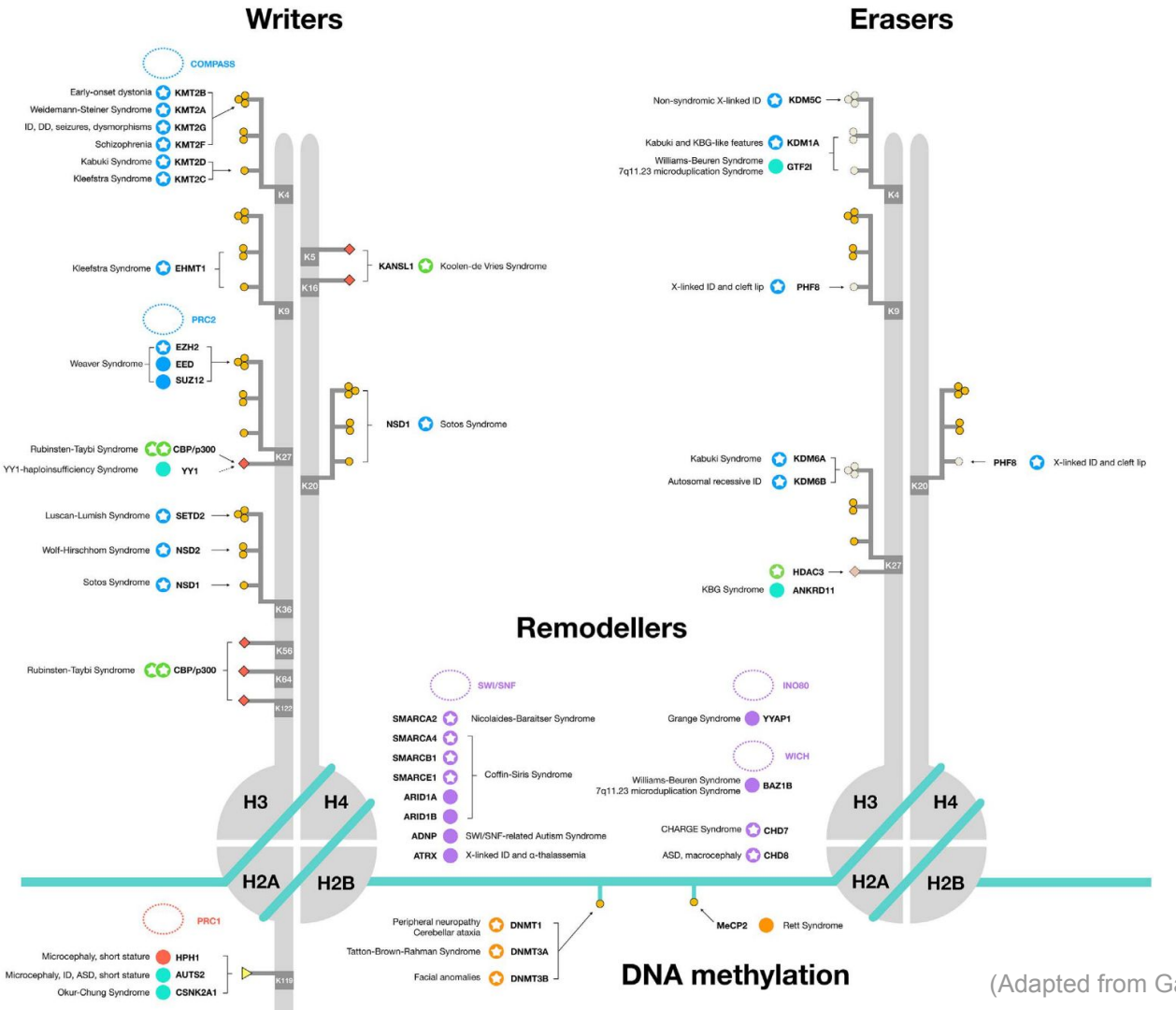
Due to the very high sparsity and noise of the data, this does not work well.
What does work well, however, is to first aggregate together groups of cells (*meta-cells*) that are highly similar, and then test correlations across meta-cells (Pliner et al., 2018; Persad et al., 2023)



# Chromatin and disease

# Autism-associated genes are enriched for chromatin-binding






Nearly all epigenetic modifier genes are associated with neurodevelopmental syndromes

"The prolonged unfolding of neurogenic potential, along with the highly vascularized nature of the developing SVZ and SGZ, have both been invoked as reasons for the vulnerability of the developing human brain (Baburamani et al., 2012)"

(Adapted from Gabriele et al. 2018 10.1016/j.pnpbp.2017.12.013)

# Chromatin and cancer

- Mutations within chromatin remodelling complexes are estimated to affect 10-20% of all cancers, typically leading to more “open” chromatin; in most of other cancers, the machinery is indirectly affected
- “A well-known characteristic of almost all tumours is global hypomethylation and concurrent abnormal hypermethylation at localized sites such as Cpg islands”

(Zhao et al., Nat Rev Cancer 2021)

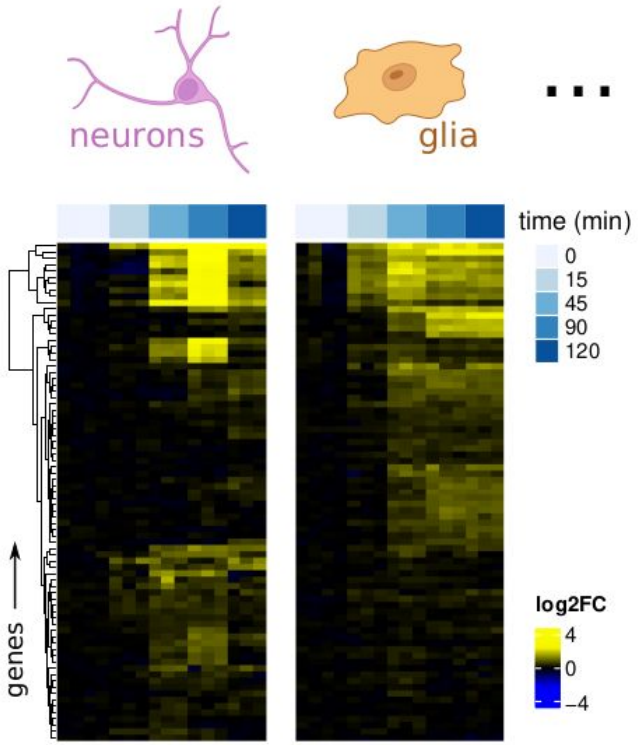
- Some cancers (e.g. infant ependymoma) don’t show relevant DNA changes, but large epigenetic alterations
- Cancerous phenotype can be induced by (mutations in) the surrounding tissue in models, in the absence of mutation of the cells themselves (see Maffini et al, J Cell Sci 2004)
- Histone acetylases (and histone deacetylase inhibitors) are having surprising success as anti-cancer drugs (they however also affect important non-chromatin pathways, such as p53 and Nfkb)

# Our lines of research

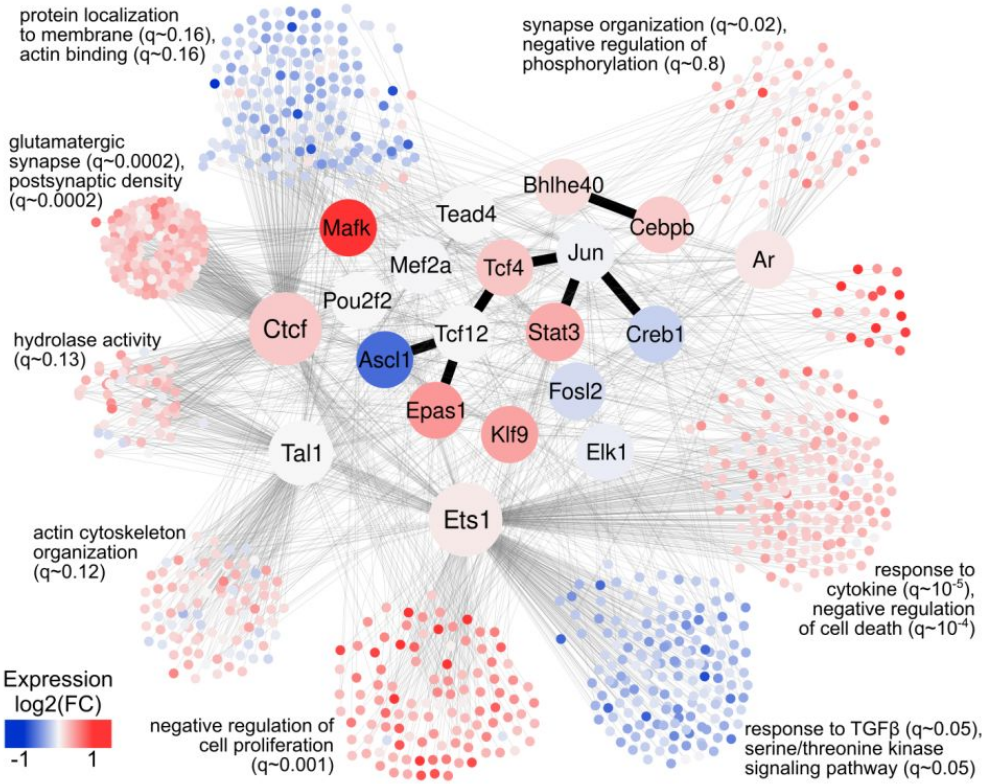
# Understanding the brain’s gene expression response to stress

- How is the (gene expression) response to stress in the brain distinguish from that to normal brain activity?
- Can we decompose this response into the contributions of different inter-cellular pathways (e.g. synaptic, hormonal)?
- How much of this response is attributable to cells simply maintaining homeostasis (e.g. metabolism, oxidative stress, etc.) in the face of intense activity?

The response of different cell types to similar stimuli is partially overlapping




Can we explain these similarities and differences in terms of combinations of TF bindings?




(von Ziegler, et al., 2022)

# Underlying computational challenges

- How can we best analyze that kind of data?
  - In particular ATAC-seq and single-cell data
- How can we get a good idea of where TF bind in different cell types, in the absence of experimental data for most TF/celltype combinations?
- How can we best make sense of distal regulatory elements and what they regulate?
- Given a gene expression signature (e.g. of a condition/disease) and transcriptional networks, how can we best infer which TFs have a differential activity?

# Course project

# Grading and expectations

- 50% of the grade is based on **weekly exercices**
  - Exercices should be **submitted via github**, by thursday noon the following week
  - The best half of the exercises will make up the grade

- 50% of the grade is based on the **project** (groups of max 3 persons)
  - The project can be either:
    - Re-producing the analyses from a publication (in a critical fashion)
    - Analyzing new data (e.g. yours or in collaboration with a group)
  - The project *must be discussed and approved in advance*
  - The expected outputs of the project are:
    - a report (e.g. ~10-15 pages) with embedded full code and figures, and including an introduction and discussion of the results
  - Deadline: before the end of the day on **July 3rd**

# Evaluation scheme for the course project

- Format and formal requirements (1/6)
  - Rendered markdown (html or pdf), proper scientific references, figure legends, etc.
- Introduction/conclusion (1/6)
  - More is not necessarily better: ask yourself what background would your fellow students need to understand your problem, analyses and observations
- Analysis
  - **Correctness**, i.e. lack of mistakes (1/6)
  - **Adaptability/creativity**, i.e. whether you could adapt (e.g. the tools seen in class) to your purposes (1/6)
  - **Appropriateness**, i.e. whether you used the right tool/visualization to address a question (1/6)
  - **Interpretation**, i.e. whether you correctly describe and interpret your figures and analysis results (1/6)
- Extra 1/6 (summing to 7 of max 6 points) for difficulty / going beyond the expectations (to not penalize those who undertook something harder)

# A few (optional) tips for preparing your project report

- It's often easier to split a project into different markdown documents for different steps
  - e.g. one that downloads the data, another processes it, another that answers specific questions of makes the figures with which you'll tell your story
  - if you want to go deeper, workflowr is a great rmarkdown organization and versioning framework

- If you're wondering how to do something in rmarkdown, consult this online book
  - See especially:
  - code chunk options
  - the section about references and bibliography
    (most reference management software, e.g. Zotero and the likes, can export references in the *bibtex* format required by rmarkdown)